Spatio-Temporal Rich Model-Based Video Steganalysis on Cross Sections of Motion Vector Planes

Kasim Tasdemir, Fatih Kurugollu, *Senior Member, IEEE*, and Sakir Sezer, *Member, IEEE*

***Abstract*— A rich model-based motion vector (MV) steganalysis benefiting from both temporal and spatial correlations of MVs is proposed in this paper. The proposed steganalysis method has a substantially superior detection accuracy than the previous methods, even the targeted ones. The improvement in detection accuracy lies in several novel approaches introduced in this paper. First, it is shown that there is a strong correlation, not only spatially but also temporally, among neighbouring MVs for longer distances. Therefore, temporal MV dependency alongside the spatial dependency is utilized for rigorous MV steganalysis. Second, unlike the filters previously used, which were heuristically designed against a specific MV steganography, a diverse set of many filters, which can capture aberrations introduced by various MV steganography methods is used. The variety and also the number of the filter kernels are substantially more than that of used in the previous ones. Besides that, filters up to fifth order are employed whereas the previous methods use at most second order filters. As a result of these, the proposed system captures various decorrelations in a wide spatio-temporal range and provides a better cover model. The proposed method is tested against the most prominent MV steganalysis and steganography methods. To the best knowledge of the authors, the experiments section has the most comprehensive tests in MV steganalysis field, including five stego and seven steganalysis methods. Test results show that the proposed method yields around 20% detection accuracy increase in low payloads and 5% in higher payloads.**

***Index Terms*— Steganalysis, steganography, video, motion vector, comparison.**

I. Introduction

STEGANOGRAPHY is an art of covert communication. The purpose of the sender, steganographer, is to hide the *existence of communication* by embedding the secret message into a carrier object and sending this innocent looking message carrier to the receiver without evoking any suspicion. The observant who has right to eavesdrop and investigate the carrier object and who is also trying to detect the *existence* of the secret message is called *steganalyzer* [1]–[3]. The message carrying and clean objects are called stego and cover respectively. Steganalyzer's job is to distinguish stego from cover. In digital steganography, the message carrier object can be of any digital medium such as image, sound, video, electronic documents, etc. Each digital medium has its own advantages. When the data size and the variety of ways to embed message are considered, video has advantages over others. Despite these advantages of video, until lately, majority of steganalysis researches has focused on images because of its popularity, ease of implementation and ease of sharing them on the Internet. However, by escalation of number of Internet users and advancements in networking infrastructures the number of videos shared online has increased almost 800% over last 6 years [4], [5]. This explosive growth of online video makes it an appealing channel for covert communication using steganography. Consequently, this fact has drawn more researchers in the area of video steganalysis.

Video codecs have more sophisticated system than does images. The secret message can be embedded into not only pixel or DCT domains but also codec specific attributes, e.g., motion vectors (MVs), macro block (MB) intra-prediction types, etc. In this study, we focus on detection of the embedded message into MV patterns using motion vector (MV) steganography methods.

Since the research on image steganalysis is more mature, a sensible approach to the video steganalysis problem would be to import and adapt the promising tools from image steganalysis side. A recent trend in image steganalysis is rich model based steganalysis. Fridrich and Kodovský [6] proposed this model for a steganalytic system against spatial image steganography. Later, Kodovský *et al.* adapted it for JPEG images [7]. The distinctive attribute of the rich model based methods is that it can capture many weak traces of steganography. Therefore, a more precise cover model can be obtained. However, Spatial Rich Model (SRM) [6] or Cartesian Calibrated JPEG Rich Model (CC-JRM) [7] methods cannot be directly applied to MV patterns because MVs are distinctly different from image pixels and this difference is rather fundamental. Firstly, source of pixel values of natural images is related to a static scene affected by illumination conditions as well as reflectance properties of objects whereas that of MVs is motion of the objects or the camera. Hence, the behaviour of the MVs are different. Secondly, MVs are *vectors* and image pixel values are scalar numbers. Thirdly, MVs have temporal information as well. Hence, any image steganalysis method requires significant modifications before it is applied to MV based steganalysis.

Manuscript received July 6, 2015; accepted April 26, 2016. Date of publication May 11, 2016; date of current version June 1, 2016. This work was supported by the Engineering and Physical Sciences Research Council through the CSIT 2 Project under Grant EP/N508664/1. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Charles Boncelet.

K. Tasdemir is with the Department of Computer Engineering, Abdullah Gül University, Kayseri 38080, Turkey (e-mail: kasim.tasdemir@agu.edu.tr).

F. Kurugollu and S. Sezer are with the Institute of Electronics, Communications and Information Technology, Queen's University Belfast, Belfast BT3 9DT U.K. (e-mail: f.kurugollu@qub.ac.uk; s.sezer@qub.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2016.2567073

In the light of the aforementioned findings, we designed a rich model based universal steganalytic system against MV steganography. The improved detection accuracy of the proposed algorithm lies in its five novel contributions:

- The temporal correlation of close and distant frames are incorporated in the system.
- Higher order filters are employed. Filters up to 5th order are used whereas previous methods use up to 2nd order filters at most.
- Rich variety and quantity of the filters (28 filter kernels) give the proposed method the ability to capture dependencies in a wide spatio-temporal range.
- The previous methods are based on a narrow assumption on MV modification method of stego algorithms such as adding and subtracting one MV value. This questionable assumption makes them rather targeted steganalysis. We do not target a specific stego algorithm.
- Previous methods used heuristic approaches, i.e., they do not benefit the theoretical or experimental advancements in image steganalysis. In this work, we adapt rich model based steganalysis, whose competence has been already confirmed for JPEG [7] and raw images [6], for use by video steganalysis.

The paper is structured as follows. The previous MV steganography and steganalysis methods are alluded briefly in Section II. The proposed MV steganalysis method is introduced and described in Section III. Section IV contains a comprehensive comparative tests of MV steganalysis and steganography methods. Section IV also has several subsections where we discuss the advantages and handicaps of the methods tested. Moreover, the reasons why the proposed method has a better detection accuracy are discussed along with its drawbacks. Besides, the best settings for the proposed method is provided through several tests in the same section. Finally, the paper is concluded in Section V.

Everywhere in the paper, capital-case boldface symbols are used for matrices and higher dimensional arrays and lowercase boldface symbols represent vectors.

The terms *spatial* and *temporal* are widely used throughout the paper but we need to clarify a possible ambiguity here. *Spatial* do not refer to pixel values. Since we deal with MVs, *spatial* should be associated with MV values belonging to the same time instant, i.e., the same frame.

To facilitate the comprehension of the paper, we use a naming convention based on the first author names instead of numbering the methods. The previous method [8] by Deng *et al.*, which uses reconstructed MVs, is named as *DengRec* and the method [9] by the same authors, which incorporates center of mass (COM), is named as *DengCom*. The other steganography and steganalysis methods [10]–[19] are named respectively as *Xu*, *Fang*, *He*, *Zhang*, *Aly*, *CaoStego*, *Su*, *Cao*, *AoSO*, *Tasdemir*, *Accordion unfolding SRM (ASRM)*.

II. Previous Methods

Currently, there are limited number of MV steganography [10]–[12], [14], [20]–[27] and MV steganalysis algorithms [8], [9], [13], [15]–[18], [28].

MV steganography algorithms can be categorized into two groups according to their MV modification strategies such as magnitude modifying and phase modifying. Magnitude modifying stego algorithms add or subtract 1 from the magnitude value of the candidate MV component/s. On the other hand, phase modifying stegonography algorithms divide cartesian coordinate system into imaginary sections each of which corresponds to bit 0 or 1. Then, the candidate MVs are rotated so that they are positioned in an appropriate section which is in agreement with the secret message bit. Although phase modifying stego algorithms evidently cause more degradations in MV patterns than magnitude modifying stego algorithms, they are, however, surprisingly more secure against previous MV steganalysis methods. The reason is that the vast majority of the previous MV steganalysis algorithms design their features by assuming that the magnitudes of message carrier MV components are slightly modified [8], [9], [13], [17], [18], [29]. For example, the most recent MV steganalysis method, AoSO [17], investigates whether the current MV is shrunk or enlarged by 1. It, therefore, fails to detect phase modifying stego algorithms.

Video has temporal information as well as spatial. In this study, it is shown that temporal dependency is as strong as spatial dependency in MVs. Therefore, in order to design a robust steganalytic system temporal dependency should also be taken into account. There are some methods which considers temporal dependency [9], [15], [19], [20], [28] and some other which does not [8], [16]–[18]. However, none of the previous methods considers temporally neighbouring frames further than one next or previous frames. It is also shown that temporal correlation of the MVs remains for temporal distances more than next five frames, depending on the content of the video. Thus, it is conjectured that considering longer distant frames will result in better accuracy in steganalysis.

Although there are also slightly related video steganography methods which do not directly modify MVs but interfere motion estimation process such as [27], they are not taken into account in this work since only the methods which specifically changes the MVs are considered.

General approach of MV steganalysis algorithms is modelling stego disturbance on MV patterns and generating features representing this disturbance. Then a machine learning classifier is trained with these features and it is used to classify a given video into stego or cover classes. Conventionally in MV steganalysis works, the disturbance due to embedding is modelled as additive independent noise on MV magnitudes. Majority of MV steganalysis algorithms are based on this assumption [8], [9], [13], [17], [18], [29]. Even if using this distribution could be effective against LSB steganography, this model cannot be used to detect phase modifying stego algorithms as the MV is moved more than one unit distance. This fact will be also demonstrated in the test section experimentally (Section IV).

Most of MV steganalysis algorithms employ high-pass spatial filters to extract the features. Earlier methods use first order spatial filters but then the following methods integrated second order spatial filters as well. Then, Ye *et al.* [28] showed that using temporal filters contributes to the accuracy which

agrees what is shown in Fig. 2. In this study, the order of filters is risen up to five in both spatial and temporal domain in order to capture the decorrelation in the underlying distribution of motion vectors due to embedding in a wide range.

Another approach to MV steganalysis problem is to obtain a cover model rather than modelling the stego message disturbance. The methods in this category [8], [16], [17] try to find the original MV by different means such as re-compressing and decompressing the video or using neighbouring MVs etc.

The very first MV steganalysis algorithm [13] and the following methods [9], [15] investigate the first order statistics of corruptions caused by LSB embedding.[1]

Cao *et al.* states that expected values of MVs of a re-compressed video are equal to that of the cover video [16]. In other words, they state that we can recover the original MVs with a high probability by decompressing and compressing a stego video. By using this assumption, their method utilises the distance between MVs before and after re-compression. However there is a problem with this assumption. In a realistic scenario the search algorithm used in the motion estimation stage is unknown to the steganalyzer. It is stated in [17] that Cao's method suffers if a different search algorithm is used in the second compression. This statement is investigated and confirmed in Section IV-D.

Another MV steganalysis method based on MV recovery has been presented in [8]. In this method, a lost MV recovery algorithm using *polynomial kernel regression* on 8 neighbouring MVs is proposed. Then this lost MV recovery method is employed for estimation of the cover MVs.

A targeted MV steganalysis for LSB steganography has been proposed by Tasdemir *et al.* [18]. It is stated that flat areas, which are common in MV patterns, are highly corrupted by LSB embedding. The statement is supported by a theoretical proof. Their features are based on the corruptions on flat areas. Despite the fact that it accurately models LSB corruption on MVs, it cannot be used against phase modifying MV steganography methods.

The most recent MV steganalysis algorithm, AoSO [17], performs a local MV search around current MV in a one-pixel wide search window. It is reported that if the new MV is found in the same position of the current MV, then it is more likely to be a cover video. It is more likely to be a stego if the optimum MV is found one-pixel close to the current MV. However, it is not clear whether half pel[2] or quarter pel resolution is taken into account. That is, if quarter pel search window width is used, then the stego MV will be $\frac{1}{4}$ pixel away from the current MV. Half pel resolution is adopted in our AoSO implementation since it is the typical setting in a real scenario. AoSO has also several other pitfalls. Even if AoSO's assumption is true, when phase modifying stego algorithms are used, MVs are always going to be the local minimum because motion estimation is performed in a large region. Therefore, AoSO's one-pixel wide search window will miss it. Another problem is with B-type MBs. These MBs have both forward and backward MVs, which points to future and past frames. When searching for a new MV, the residual error is calculated by extracting current MB from the average of future MB and the past MB. Therefore their aggregate residual error is minimum rather than individual residual errors. Because of that, intuitively the optimum forward MV and backward MV should have been searched simultaneously in AoSO [17].

III. Proposed Spatio-Temporal Rich Model (STRM)

In this section, first a new term, motion vector plane, is defined and then the proposed Spatial-Temporal Rich Model (STRM) for video steganalysis is presented in Section III-B by pointing out the contributions over SRM and other previous MV based steganalysis methods.

A. Motion Vector Planes

Each MV has x and y components. If the macroblock is of type B, its corresponding MVs have four components, i.e., x and y for both backward and forward predictions. MV patterns of a B frame is distinctly different than that of P frame. Therefore, MVs should be grouped according to their frame types. This technique was used before by [9] and [15] as well. In our method, we group *components* of MVs together according to their frame type and carry out feature extraction on these groups individually. Or more precisely, same type of directional components (x or y) of MVs of the frames with same frame type (B or P) and same prediction type (forward or backward predicted) are grouped together. It would be useful to introduce a new term here. Each MV matrices of **a frame** with same prediction type and same component is named as *MV plane*. For example, a typical B frame would have 4 MV planes (forward predicted x and y components, backward predicted x and y components). Let $V \in \{-W, \ldots, W\}$ represent a MV component where W is the search window width in motion estimation (ME). Then, all sets of MVs in a video is denoted as $\mathbf{MV} = (MV_{i,j,k}) \subseteq \{\cup_{c\in\{x,y\},d\in\{f,b\}} V^{(c,d)}\}^{M\times N}$ where c denotes the direction of the component (x or y), d stands for the prediction direction (f and b are abbreviations for *forward prediction* and *backward prediction*). Spatial and temporal coordinate of the corresponding MB is represented by (i, j) and k respectively. The symbols M and N are used for row and column size of 2D macroblock array in a frame. (For example, a CIF video is of size 322×288 pixels. Then, it would have $M = (288/16) = 18$ and $N = (322/16) = 22$, if the MB size[3] is 16×16). MV plane is 2D matrix slice of **MV** at a specific frame, type and coordinate, $\mathbf{V}_{k_1}^{(c_1,d_1)} = (V_{i,j,k=k_1}^{c=c_1,d=d_1}) \in \{-W, \ldots, W\}^{M\times N}$. Formal definition of a MV plane block is

$$\mathbf{V}_{\mathbf{k}}^{(c_1,d_1)}, \mathbf{k} = (n-2, n-1, n, n+1, n+2) \qquad (1)$$

where n is an arbitrary number corresponding to the order of frame of the current MV plane and $\mathbf{V}_{\mathbf{k}}^{(c_1,d_1)} \in \{-W, \ldots, W\}^{M\times N\times5}$. Feature extractions will be carried out on the MV plane blocks as explained in Section III. The reason why five MV planes are taken into account will be explained at the end of this subsection.

[1] In this brief review, only the related part of the methods are introduced. The readers are referred to the original publications for further details.

[2] *pel* is a technical term used in video coding standards. It is a measure for the pixel distance.

[3] MB size can vary according to the codec used. MPEG-1/2 is used in the examples.

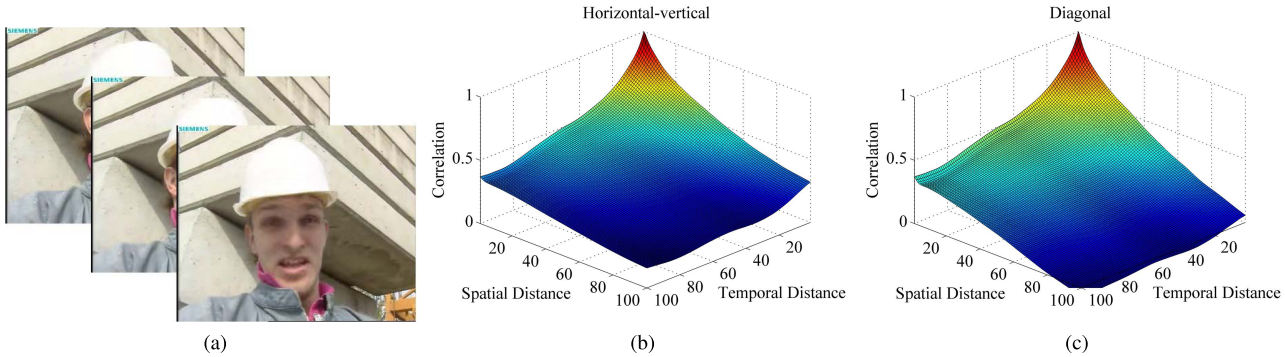

Fig. 1. Temporally and spatially neighbouring **pixel** correlation of consequent 100 frames of a video. (a) Three frames of video *foreman* (25 fps). Correlation of (b) Horizontal and vertical, (c) diagonal and minor-diagonal neighbouring pixels.

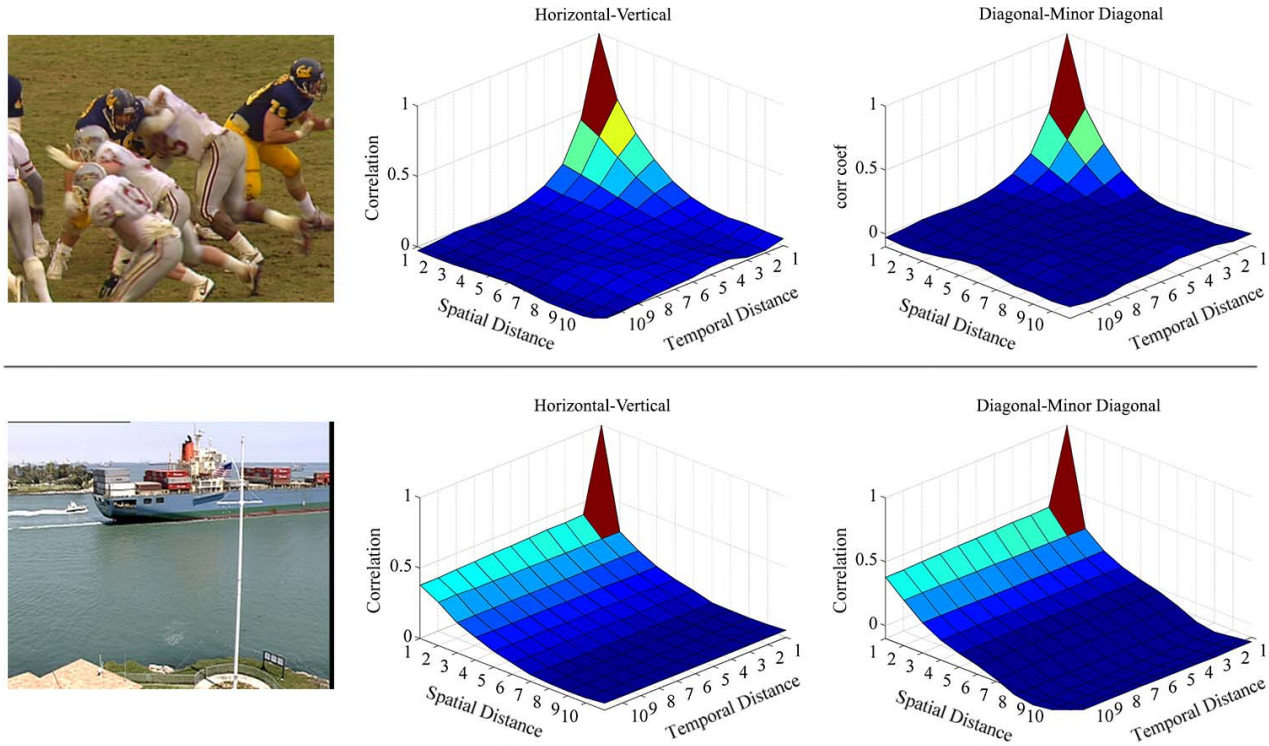

Fig. 2. Correlations of spatially and temporally neighbouring **MVs** of a slow and fast video. The first column shows Screenshots of a fast moving video *football* (25 fps) (above) and a slow moving video *container* (25 fps) (below). The second column depicts the Pearson correlation of horizontally and vertically neighbouring MVs while the third Column illustrates that of diagonally and minor diagonally neighbouring MVs.

MVs have different characteristics than image pixels because of two major reasons. First, their sources are distinct. MVs are associated with the motion of the objects or the camera where image pixels are effected from intensity and colour of the reflected light. Secondly, the number of pixels in a frame is much higher than that of MVs. For example, there are typically 256 times more pixels than MVs in a MPEG1/2 video. Or more precisely, (number of pixels)/(number of MVs) ratio in MPEG1/2 video is in 128-256 range. This ratio is between 8-256 in a H.264, 8-4096 in a HEVC video. These differences give both advantages and disadvantages over images in terms of steganalysis. The disadvantage of MVs is that they have weaker correlation compared with pixel values. Fig. 1 shows that the correlation of temporally and spatially neighbouring pixels remains over 0.75 for around 10 pixel distances whereas that of MVs, see Fig. 2, quickly drops below 0.6 after only one MB distance. Therefore, a cover model of MVs based on neighbouring MVs would not be as accurate as for that of image pixels. Hence, detection rates obtained in image steganalysis is practically unattainable for MV steganalysis if only spatial dependency in a frame is exploited. Unlike images, however, MVs have temporal correlation. Thus temporal dependency can be used in order to reveal any aberrations in MVs. In order to decide on the temporal and spatial coverage of the system to be designed, first, correlations of MVs of two videos for different spatial and temporal neighbouring distances are investigated as illustrated in Fig. 2. A fast moving video *football* (above) and a

slow moving video *container* (below) are used in the test to highlight the difference. Axes of the plots on second and third row correspond to temporal and spatial neighbouring distance of compared MVs. The vertical axis is Pearson correlation coefficient. Pearson correlation coefficient is calculated for up to six spatial and five temporal MV neighbouring distance. It is evident from the Fig. 2 that there exist a temporal correlation as well as a spatial one in MVs of a video. The strength of the correlation is highly related with both spatio-temporal distance and content of the video. If the video has steadily slow moving objects (see Fig. 2b), its MVs have smaller values and are more susceptible to be suppressed by noise. Therefore, spatial correlation falls down quicker in a slow video. However, temporal correlation retains its strength for a longer period of time because the moving object, which is the source of the salient MVs, remains its position in the scene longer. On the contrary, MV correlation of a fast moving video (see Fig. 2a) exhibits opposite behaviour where spatial correlation is stronger than temporal one. In this case, lower temporal correlation is expected because objects change their position quicker. The spatial correlation comes from larger MV magnitudes and the bulk moving regions which is common in fast videos. Fig. 2 also shows that major, minor diagonal or horizontal, vertical neighbouring MVs have similar correlations. In short, incorporating both spatial and temporal dependencies in a MV steganalysis for around five neighbouring distances is essential for a reliable detection because the correlation between MVs of the MBs closer than five spatio-temporal MB distance remains strong even in a fast video.

B. Proposed Method

In the proposed method, features are derived from the predictions of MVs from their neighbourhood. More formally the residual at i^{th} row and j^{th} column $R_{ij} \in \mathbb{R}^{M\times N}$ is defined as:

$$R_{ij} = \hat{V}_{ij}(V^{(c_1,d_1)}_{i\neq x, j\neq y, k_1}) - cV^{(c_1,d_1)}_{i=x, j=y, k_1} \tag{2}$$

where $\hat{V}_{ij}(.)$ is a predictor of $cV^{(c_1,d_1)}_{i,j,k_1}$ using neighbourhood of $V^{(c_1,d_1)}_{i,j,k_1}$ and $c \in \mathbb{N}$ is residual order. Before calculating the co-occurrence matrix, residuals $\mathbf{R}$ need to be truncated by a quantization step $q > 0$:

$$R_{ij} \leftarrow \text{trunc}_T\left(\text{round}\left(\frac{R_{ij}}{q}\right)\right) \tag{3}$$

The quantization reduces the range of residuals allowing co-occurrence matrices describe them within a small range $[-T, \ldots T]$. Calculation of co-occurrence matrix is carried out as described in [6].

High-pass filters up to 5^{th} order are employed. Not only their horizontally and vertically rotated versions but also their diagonally rotated versions are used. As shown in Fig. 2 even diagonal correlation of MVs are slightly less stronger than horizontal and vertical ones, it is still strong enough to incorporate in video steganalysis. Therefore, we also bring diagonal versions of the filters into the proposed method as in [6].

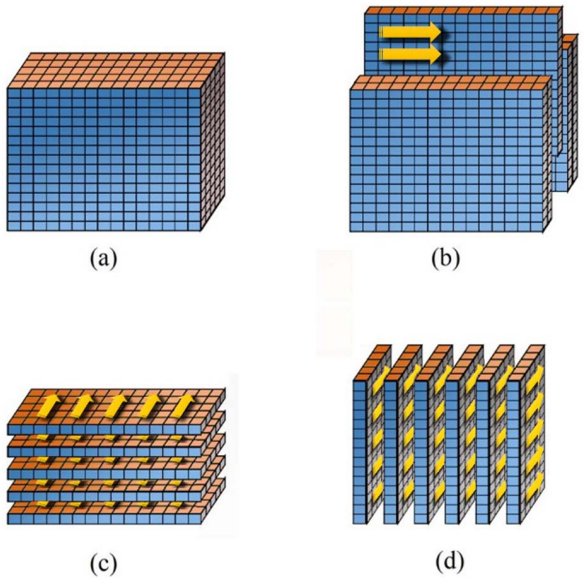

Fig. 3. Utilizing rich filters on both spatial and temporal domain in the proposed method. (a) MV plane block, (b) current MV block, which is the one in the middle, (c) horizontal-temporal cross section of MV block, and (d) vertical-temporal cross section of MV block.

SRM is only applicable to 2D data. However, MV information of a video has at least 3 dimensions considering the time, or the temporal, dimension. This problem is alleviated by using *accordion unfolding* transformation proposed in [19]. In this study, instead of transforming 3D data into 2D as in [19], rich filters are applied on horizontal cross sections, vertical cross sections and spatial cross section (current MV plane) of MV plane blocks as demonstrated in Fig. 3. First, a MV plane block, which is formally defined in Eq. 1, is formed by concatenation of consequent MV planes. The yielding MV plane block for a CIF sized MPEG2 video is illustrated in Fig. 3a. Then, the MV plane block in Fig. 3a is sliced in spatial (Fig. 3b), horizontal-temporal (Fig. 3c) and vertical-temporal (Fig. 3d) directions. Each filter is applied to both spatial MV plane, which is the current MV plane, the horizontal-temporal cross sections, and vertical-temporal cross sections. In the figure, the yellow arrows show the surfaces that the filters are applied to.

The horizontal-temporal slicing operation is explained in Algorithm 1. It takes six arguments: $\overline{\mathbf{V}}$, k, c, d, q and T. The arguments k, c and d respectively stand for current MV plane number in the group, component type (x or y), and prediction type (forward or backward). $\overline{\mathbf{V}}$ is a five dimensional matrix which contains all MV information of either all B or all P frames of the video (As stated before, B and P frames are grouped into two separate sets). The first two dimensions of $\overline{\mathbf{V}}$ stands for the MV position (row and column number). The third dimension shows the frame number in the group. The fourth dimension of $\overline{\mathbf{V}}$ is c, which indicates if the element is x or y component of the MV. Similarly, the fifth index, d, shows whether it is forward or backward prediction. MVs which has same prediction direction and which are the same type of components (x or y) are stored in a variable $\widetilde{\mathbf{V}}$. It is a 3D MV plane block, depicted in Fig. 3a, which has

Algorithm 1 Pseudocode for horizontal-temporal slicing operation and co-occurrence matrix extraction.

procedure HORIZONTAL-TEMPORAL SLICING($\overline{\mathbf{V}}$, k, c, d, q, T)
  $\widetilde{\mathbf{V}} \leftarrow \overline{\mathbf{V}}[:,:,:,c,d]$ ▷ MV plane block is formed.
  for i=1 to M **do**
    for j=1 to N **do**
      HT$[j, \text{end+1:end+5}] \leftarrow \widetilde{\mathbf{V}}[i, j, \text{k-2:k+2}]$
    end for
    Residuals ← Apply Rich Filters to **HT**.
  end for
  Residuals ← Quantize Residuals with q
  Residuals ← Truncate Residuals with T
  return Calculate Co-occurrence matrices from the Residuals
end procedure

all either x or y components of MVs of same prediction direction. The other two arguments, q and T, are the required parameters of quantisation and truncation stages. First, first rows of two previous, current and two next MV planes are concatenated to form a 2D matrix, which is named **HT**. Then, residuals are extracted from **HT**. This process is repeated for each row of the MV plane block. Then, the residuals are quantised with a scalar q. Subsequently, the quantised residuals are truncated with parameter T. Finally, features, the co-occurrence matrices, are calculated from the quantised and truncated residuals. Vertical-Temporal slicing and feature extraction is performed similarly.

The spatial slice of a MV plane block has horizontal and vertical statistical symmetry as in images. That is, we expect to have similar statistics of MVs if the slice is rotated 90°. However, unlike Fig. 3b, the temporal cross sections Fig. 3c and 3d do not have statistical symmetry in all 4 directions because a spatial slice contains intra-dependencies of MVs in a frame whereas temporal cross sections exhibits inter-dependencies of MVs among frames. They have symmetry from left to right or top to bottom individually. Hence, vertical and horizontal co-occurrence matrices cannot be added for temporal cross sections; the resulting feature vector size increases in temporal cross sections. Therefore, spatial and temporal cross sections yield different size of feature sets.

Individual and collective performance of feature vectors extracted from spatial, horizontal-temporal (HT), vertical-temporal (VT) cross sections are shown in Fig. 4. Fig. 4 is obtained by applying the proposed algorithm only with spatial, VT and HT features to all of the data set used in Section IV and then averaging the results. It is evident from Fig. 4 that temporal data contributes more than spatial data but the collection of all features provides the best results.

1) Structure of the Feature Vector: For an easier interpretation of how the filters given in [6] are adopted, one of these filters is taken as an example and feature extraction from MV planes is demonstrated in this section. Let's take the following second order minmax filter as an example:

$$R_{ij} = min\{X_{i,j-1} - 2X_{i,j} + X_{i,j+1}, X_{i-1,j} - 2X_{i,j} + X_{i+1,j}\} \tag{4}$$

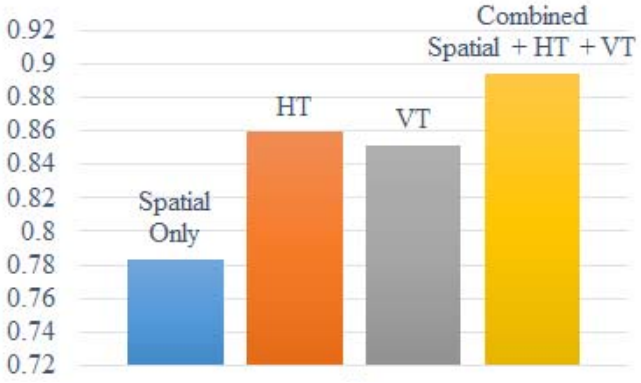

Fig. 4. Effect of sub-feature sets given in Eq. 8. Features obtained from HT and VT cross sections give a better model than does spatial features. Combining the three sub-feature set gives the best accuracy.

where $X_{i,j}$ is the image pixel value at ith row, jth column. As this is horizontal-vertical (HV) symmetrical filter, there is no need to calculate the residuals of 90° rotated version of the this filter. This filter is applied to spatial (HV cross section) (Fig. 3b), horizontal-temporal cross section (Fig. 3c), vertical-temporal cross section (Fig. 3d) of the current MV plane block. Structures of these filters are as follows:

$$\begin{aligned} Rspatial^{c,d}_{i,j,k} &= min\{V^{c,d}_{i,j-1,k} - 2V^{c,d}_{i,j,k} + V^{c,d}_{i,j+1,k}, \\ &\qquad V^{c,d}_{i-1,j,k} - 2V^{c,d}_{i,j,k} + V^{c,d}_{i+1,j,k}\}, \\ i \in \{2,\ldots,M-1\}, &\quad j \in \{2,\ldots,N-1\},\ k \in \{n\} \end{aligned} \tag{5}$$

$$\begin{aligned} Rht^{c,d}_{i,j,k} &= min\{V^{c,d}_{i,j-1,k} - 2V^{c,d}_{i,j,k} + V^{c,d}_{i,j+1,k}, \\ &\qquad V^{c,d}_{i,j,k-1} - 2V^{c,d}_{i,j,k} + V^{c,d}_{i,j,k+1}\}, \\ i \in \{1,\ldots,M\}, &\quad j \in \{2,\ldots,N-1\},\ k \in \{n-1,n,n+1\} \end{aligned} \tag{6}$$

$$\begin{aligned} Rvt^{c,d}_{i,j,k} &= min\{V^{c,d}_{i-1,j,k} - 2V^{c,d}_{i,j,k} + V^{c,d}_{i+1,j,k}, \\ &\qquad V^{c,d}_{i,j,k-1} - 2V^{c,d}_{i,j,k} + V^{c,d}_{i,j,k+1}\} \\ i \in \{2,\ldots,M-1\}, &\quad j \in \{1,\ldots,N\},\ k \in \{n-1,n,n+1\} \end{aligned} \tag{7}$$

where n is the current MV plane number, $c \in \{x, y\}$ and $d \in \{\text{forward predicted}, \text{backward predicted}\}$. Residual subscripts *spatial*, *ht* and *vt* indicates that these residuals are obtained from the cross sections spatial, horizontal-temporal and vertical-temporal. Then, the residuals are quantized as in Eq. 3. It is found that quantizing with various quantization values does not noticeably improve accuracy (See Table I). Thus, we only take $q = 1$ in order to reduce the feature size.

In our proposed method, these filters are not only applied to spatial (horizontal-vertical) cross section of MV plane block, but also to its horizontal-temporal and vertical-temporal cross sections. HT and VT co-occurrence matrices requires different dimension reduction approach than that is used for images because the residuals have different statistics in temporal and spatial dimensions. It is assumed that natural images would have the same statistics if the image is rotated by 90°. However in our case, this assumption only can be used for spatial

TABLE I

TESTING THE ACCURACY DIFFERENCES WHEN DIFFERENT QUANTIZATION AND CO-OCCURRENCE MATRIX SIZES ARE USED IN VARIOUS PAYLOAD RANGES. THE BEST SETTINGS ARE FOUND TO BE $q = 1$ AND $T = 2$

Payload range	$A_{q=(1,1.5,2)} - A_{q=1}$	$A_{T=2} - A_{T=1}$
0.0 - 0.1	6.95	87.93
0.1 - 0.2	5.05	36.33
0.2 - 0.3	-0.92	44.85
0.3 - 0.4	0.36	20.63
0.4 - 0.5	-7.59	16.48
0.5 - 0.6	-8.25	29.32
0.6 - 0.7	8.56	19.09
0.7 - 0.8	2.50	19.31
0.8 - 0.9	-3.35	14.97
0.9 - 1.0	5.67	12.83

residuals such as eq. 5. There is only horizontal, vertical and temporal symmetry individually in the statistics of a MV plane block. Hence, it is necessary to take horizontal and vertical co-occurrence matrices of HT and VT cross sections into account individually. That is, there are two co-occurrence matrices for each filter whether or not it is HV symmetrical. As a result the total feature set size is $12753+2\times\big((7\times2)\times169+(21\times2)\times325\big) = 12753+16016+16016 = 44785$ unlike in SRM.

Our final feature vector is the collection of spatial, HT and VT sub-feature vectors as follows:

$$\overbrace{F_{\text{combined}}}^{44785}: \quad (\overbrace{F_{\text{spatial}}}^{12753}, \overbrace{F_{\text{ht}}}^{16016}, \overbrace{F_{\text{vt}}}^{16016}). \tag{8}$$

Individual dimensions of the feature vectors in Eq. 8 are given above each one. Contribution of each sub-feature vector and the final feature vector F_{combined} is shown in Fig. 4. We have employed all the filters given in [6] to spatial, HT and VT cross sections of MV plane blocks but these filters are not explicitly listed here for brevity.

IV. COMPARATIVE TESTS

The purpose of this section is twofold. The first one is to demonstrate the contribution of the proposed method over previous adversary methods and to investigate its performance individually against current MV steganography methods. The second one is to give a clear cross comparison by investigating both the accuracy of MV steganalysis algorithms and also the security of MV steganography algorithms in various payloads. For an objective comparison, the most recent MV steganalysis method AoSO [17] and other 3 steganalysis methods, namely, DengRec [8], DengCom [9], Su *et al.* [15] are implemented. The most prominent MV steganography methods Xu *et al.* [10], Fang and Chang [11], He and Luo [12], Pan *et al.* [23], and Aly [24] are also implemented for message embedding by means of an open source library [30]. Some steganalysis methods mentioned above are only applicable to P frames only. If such methods were being tested, IPPP GOP structure was used. IBPBI GOP structure is used for the rest. Except GOP structures, 100 unique CIF sized videos each containing 100 frames are encoded with the same settings. Full parameter settings are as follows:

- Size: 352 × 88 (CIF) Coloured
- Bitrate: 1152000 bit/s
- Framerate: 25 fps
- Mode: Progressive
- Chroma sub-sampling: 4:2:0
- Stream type: ISO/IEC 11172-2
- Aspect ratio: CCIR601 625 line
- Video format: PAL
- Intra DC precision: 8 bit
- P frame W: Horizontal: 13 Vertical: 13
- B frame W: Horizontal: 8 Vertical: 8
- Motion search algorithm: Exhausted (full)
- Motion search resolution: Half pel

TABLE II

TEST SETTINGS FOR STEGANOGRAPHY ALGORITHMS USED IN THE TESTS

Name	Threshold	Number of Regions
Aly	PEF min: 15db max: 60db	-
Xu	5	-
Pan	5	16
He	5	16
Fang	5	8

The final data set was comprised of total 700 stego and cover videos which makes 700000 frames on total (550000 P type MV plane, 1000000 B type MV plane). Settings of the steganography methods used in our tests are given in Table II. Prediction Error Frame (PEF) limits for Aly's embedding method are set to 15db for minimum and to 60db for maximum. Thresholds of other embedding methods are set to 5 and number of regions are set to 16, 16 and 8 for Pan, He and Fang respectively

In steganography, payload is the relative amount of information embedded into the carrier object. When the amount of the information in the object is higher, more degradation in it is expected and steganalysis is expected to be easier. Conventionally, payload can be determined by the sender in image steganography methods. However, all MV steganography algorithms allow user to alter the payload by only choosing an appropriate MV threshold rather than payload itself directly. Thus, a stego video data set with a predefined payload cannot be generated in a realistic scenario. Therefore, we carried out tests for various thresholds and split the test set with respect to resulting payload range instead of modifying the original stego algorithms to restrict them into a payload range. Thus, different payload ranges include different amount of data set, e.g., in our test set, payload range 0.0-0.1 has substantially more samples than 0.5-0.6. This results in a slight drop in detection accuracies of the tested steganalysis algorithms especially in payload range 0.5-0.7.

The test results in the plots are left empty if there is less than 10 samples in that payload range (For example, DengCom has no test results for payload range 0.8-0.9). Tests are carried out for ten different ranges of payloads. Nevertheless, the conventional meaning of the term *payload* is slightly abused here because of the reasons elaborated on presently.

The feature vectors of some steganalysis algorithms are extracted from a *unit* containing a group of frames. For example, a unit contains 6 consequent same type of frames

TABLE III
OPTIMUM SETTINGS FOR PROPOSED STRM OBTAINED FROM TRAINING AGAINST 5 MV STEGO METHODS IN VARIOUS PAYLOAD RANGES ARE GIVEN

	Payload	0.0 - 0.1	0.1 - 0.2	0.2 - 0.3	0.3 - 0.4	0.4 - 0.5	0.5 - 0.6	0.6 - 0.7	0.7 - 0.8	0.8 - 0.9	0.9 - 1.0
Aly	OOB (10^{-3})	145.5	110.0	104.1	72.5	65.4	47.8	33.9	28.5	31.0	7.6
	D_{sub}	1800	1200	800	850	775	775	600	400	400	1000
	L	220	179	109	230	240	235	168	147	111	104
Fang	OOB (10^{-3})	258.4	79.2	61.9	40.5	55.9	109.5	50.3	15.2	18.2	3.8
	D_{sub}	2000	800	800	500	400	250	400	300	400	300
	L	197	145	359	391	500	500	319	339	152	74
He	OOB (10^{-3})	321.5	155.8	165.6	146.4	121.5	235.8	180.3	114.2	74.0	41.0
	D_{sub}	1600	925	325	325	219	250	263	300	200	338
	L	179	360	471	500	320	500	500	500	289	500
Pan	OOB (10^{-3})	325.1	159.6	144.1	132.4	118.8	252.4	156.3	87.1	47.6	41.0
	D_{sub}	1800	925	600	300	350	262	250	325	350	300
	L	168	255	500	500	500	500	500	500	290	304
Xu	OOB (10^{-3})	264.7	149.3	131.1	83.7	65.1	70.8	44.4	41.2	25.3	26.3
	D_{sub}	2400	950	487	450	300	225	200	200	575	300
	L	207	325	500	500	273	500	500	145	494	478

in [15] where a unit contains 1 MV plane in [19]. This makes it impossible to carry out a pairwise comparison because of two reasons. Firstly, when a unit is classified as stego by a steganalysis method, it does not say anything about individual frames in that unit. Secondly, when comparing them in various payload ranges there would be many gaps in the higher ranges, since it is unlikely to have a unit with full embedding rate, where all consequent 15 MV planes are fully embedded. In order to overcome this problem, units are sorted with respect to total amount of message bits they are carrying. We consider the unit with the highest amount of message bits as the unit with full payload even if it is not in the literal sense. Then, the data set is split into 10 payload ranges starting from 0 to full payload. Hence, the full payload in Fig. 5 means the maximum payload that could be embedded in a unit of frame/frames in the tests.

Test videos are generated using raw image sequences [31]. Since some of the steganalysis methods works only for IPPP or IBPB GOP structures, each image set is encoded two times so that each cover video has its both IPPP and IBPB GOP versions. Five stego algorithms are included in the tests. Hence, each cover video has its corresponding stego versions. A randomly generated secret message bit sequence with independent and identical distribution is embedded to each stego video. Then, MV planes are extracted from the video set and sorted by their payload. There is always a cover MV plane corresponding to each stego MV plane in a payload range. Training and test are performed for every payload range individually in order to reveal the precise detection accuracy of the MV steganalysis methods with respect to the amount of the message embedded in the stego video. Half of the stego videos in a payload range are used for training the steganalytic systems and the other half is used for testing. The videos with GOP sequence IPPP or IBPB are managed separately in trainings and tests. When a steganalysis method is trained with a stego and a cover video set of IPPP GOP structure, it is also tested with a video set of same type of GOP structure, IPPP. IBPB is used otherwise. Then the results are averaged. All settings of training and tests are applied as explained in the related publications.

In order to have a fair comparison of the MV steganalysis methods [8], [9], [15], [19] with respect to various payload ranges, they are trained half of the samples in that payload range which are randomly selected. We trained our proposed STRM against every payload range and every steganographic method as well. We have used an ensemble classifier as it is more convenient for larger data sets [6], [7], [32], [33]. Table III shows the optimum settings for our method found in the training stage. Out of bag error (OOB) shows the error obtained from cross validation during the training. L stands for optimum number of learners and D_{sub} shows optimum subspace dimension for that L. For a detailed descriptions see [6]. The table shows that the system has less training error in higher payloads. As we expect, OOB and D_{sub} decreases as the payload increases for any steganographic method. When the payload is high, the expected corruption in a MV plane is also high. This corruption reduces the chances of a stego evading from the ensemble classifier. We should also point out that the computational bundle is lower for higher payloads because corresponding number of learners and the subspace dimensionality is remarkably lower in contrast to other payloads.

All test results are shown in Fig. 5. The tests are repeated for 10 times. Thus, each data point in Fig. 5 is the average of 10 train-test results. Each plot in Fig. 5 shows the accuracy vs payload plots of all steganalysis algorithms against every steganography algorithm given in the title of the graph. Accuracy P_A is calculated as:

$$P_A = 1 - (P_F + P_M) \tag{9}$$

where P_F and P_M are experimental probability of false detection and miss detection respectively. We tested our proposed spatial only features and spatial + temporal features combined (STRM) in order to observe the contribution of temporal features. The tests also include our previous method accordion unfolding ASRM [19], which is the closest rival of proposed STRM.

In short, Fig. 5 shows that proposed STRM method has an outstanding accuracy in contrast to previous adversary methods against any steganography methods tested. The test

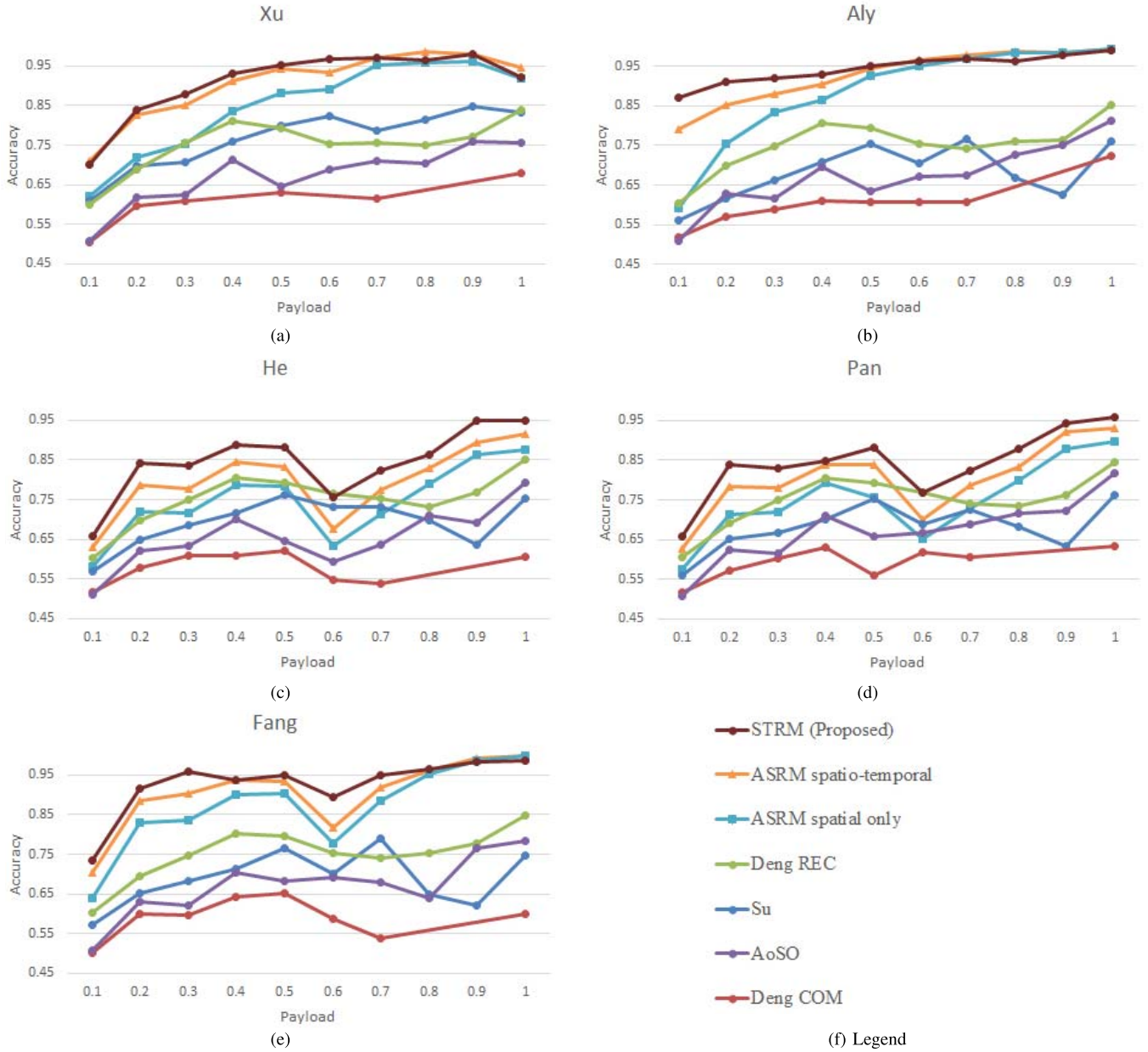

Fig. 5. Seven steganalysis methods are tested against MV steganography methods: (a) Xu, (b) Aly, (c) He, (d) Pan, (e) Fang.

results given in Fig. 5 merit an elaborate discussion. More detailed discussions are given in the following subsections.

A. Performance of the Proposed STRM Steganalysis Scheme

As it is shown in Fig. 5, overall accuracy of rich model based steganalysis methods (ASRM, STRM) are higher than the others. However, the gap between the accuracies of previous and rich model based methods differs according to used stego method in tests and the embedding ratio range. Rich model based methods have substantially better detection accuracy against LSB based steganographic methods [10], [24] and the phase modifying method, Fang and Chang [11], in any payload range. MV phase modifying methods [12], [23] is less detectable by rich model based steganalysis and others in mid and lower payload ranges. Actually, rich model which use only spatial features falls behind Deng's reconstruction based steganalysis against He and Luo [12] and Pan *et al.* [23] for lower payloads. Nevertheless, the proposed STRM algorithm succeeds both spatial only rich model and our previous Accordion Unfolding SRM (ASRM) method in every test. The major reason for this is the larger spatial and temporal coverage of the proposed method. Moreover, unlike ASRM, STRM can capture both horizontal and vertical correlations in same manner. In ASRM, MV plane block is accordion unfolded either in vertical or horizontal direction.

There is, however, a slight drop in payload $(0.5 - 0.6)$ range. This aberration in the graph is due to underfitting of the machine learning stage, which is a result of having less number of samples in that range (it was discussed in Section IV). The accuracy of rich model based steganalysis

methods goes beyond 90% after about 0.2 embedding rate against Aly, Xu and Fang. This shows that if the payload is higher than 0.2, the proposed system can reliably detect LSB based steganographic methods as well as phase modifying based ones which divides the region into less than 8 sections.

If only spatial SRM and spatio-temporal rich model are considered in Fig. 5, it is noticed that there is an ample amount of increase in accuracy thanks to the temporal features. Temporal features give extra around 5% accuracy in mid and high payloads, 20% in low payloads against Aly and Xu. These results conform with our investigation on the correlation between MV plane of current frame and that of temporally neighbouring frames depicted in Fig. 2.

B. Comparison of Motion Vector Steganalysis Methods

Results show that STRM and ASRM are better in terms of overall classification accuracy. However, this order changes for some specific cases. For instance, when the payload is greater than 0.5, Su's method has better accuracy than AoSO against Xu. On the other hand, DengRec acquires the top place only at payload range $(0.5-0.6)$ against Pan and He. These fluctuations comes from the employed machine learning algorithm and settings. As we stated before the number of samples vary in payload ranges. Some methods, especially the proposed STRM method, are more prone to underfitting problem for insufficient number of samples. For example, DengCom extract features from 15 frames. Thus, its data set size is 15 times smaller than AoSO's. Thus, DengCom requires more sample frames than AoSO does for a better training. As seen in Fig. 5c,d,e, insufficient number of samples at payload range $(0.5-0.6)$ cause a slanting drop in accuracy for our method as well.

Surprisingly, AoSO do not perform better than previous steganalysis methods against Xu and Aly. One possible reason is that half pel resolution is used in motion estimation stage of tests, which is more common than full pel resolution in a real scenario. Locally optimal MV search is bounded to a 3×3 half pel sized search box in AoSO method. Not surprisingly, AoSO fails against phase modifying stego methods because it is not designed to catch such aberrations in MV patterns. It is only supposed to work if the MV has not been changed more than 1 pel. The feature vector depends on only PEF by reorientation of a MV only by 1 pel. Nevertheless, Pan, He and Fang methods rotate the MV much further than 1 pel distance. Even the extensive modification makes these stego methods conducive to be easily detected, the limitation of the features of AoSO precludes it from discerning the abnormality. Threfore, AoSO cannot be used against phase based MV steganography algorithms.

DengRec, which uses reconstructed MVs to form a cover model, is the most competitive accuracy plot among the other methods. Nevertheless, it should be noted that only full search is used in motion estimation stage of construction and reconstruction of the videos. As reported in [17], the accuracy of video steganalysis methods relying on reconstructed MVs drastically declines if a different MV search technique is used in ME stage. This is also tested and shown in Section IV-D.

C. Comparison of Motion Vector Steganography Methods

The first finding that the tests reveal about stego methods is the superiority of MV phase based stego systems over MV magnitude based ones. As it is evident from Fig. 5a and 5b, Aly's and Xu's methods are the most vulnerable MV stego methods among others. Instead of the phases of MVs, they alter the magnitudes of MVs to embed the secret message. The MV phase altering methods, He and Pan, divide the cartesian coordinate into 16 regions where Fang divides it into 8 regions. Fig.5c,d,e shows that 16 regions are more secure than 8 regions. If the number of regions is 8, STRM can detect these methods with over a 90% accuracy after payload 0.2 as seen in Fig. 5e. It should be pointed out that a steganalytic system could be designed to reveal the corruptions in MV phase patterns but none of the current adversary methods takes MV phases into account.

The difference of Pan from He is that it uses coset syndrome coding to reduce the number of MV changes. However, coset syndrome coding did not provide a considerable extra security for Pan as they are virtually same in terms of security. The performance of both methods can be compared in Fig. 5c,d.

The most degradation in MV patterns is caused by Aly's algorithm because it alters the zero MVs as well as nonzero ones. Typically, MV patterns have wide clusters of zero MVs. Distortions in these regions can readily detected. Another drawback of Aly is when a MB is chosen as a candidate, all corresponding MVs are altered, no matter what their magnitudes are. When a message is embedded into a B frame, all four MV planes are effected at the same degree. Occasionally, some MV planes of cover videos are zero matrices. After message embedding using Aly, they are highly corrupted and makes the video vulnerable to detection.

D. Comparison to Reconstructed Motion Vectors Based Steganalysis

Cao *et al.*'s steganalysis method differs from others. Stego video is decompressed and compressed back again in order to retrieve the original MVs. However, there is a problem with this approach. Motion estimation used in the first compression is not known to the steganalyzer. If the second motion estimation method is the same, recovered MVs are most likely to be same as untampered ones. It is stated in [17] that if a different motion estimation method used in second compression, the accuracy of Cao *et al.*'s steganalysis method deteriorate sharply. Thus, the proposed STRM is compared with Cao *et al.*'s separate from the tests given in Fig. 5. First, both methods are tested when the first motion estimation (ME1) is as same as the second one (ME2). Four different ME methods are incorporated in the tests: Enhanced Predictive Zonal Search (EPZS) [34], Diamond Search (DIA) [35], Hexagon-based Search (HEX) [36] and Exhaustive or Full Search (ESA).

The same raw image sequence used in the previous test is compressed using various motion estimation methods (ME1) and a random message is embedded into them using Xu, Aly, Fang, He and Pan steganography methods using an open source library [37]. Then, the video set is decompressed

TABLE IV

DETECTION ACCURACIES OF CAO AND STRM ARE COMPARED WHEN ME1 = ME2 = EPZS AND ME1 = ME2 = DIA

		EPZS, EPZS		DIA, DIA	
	Payload Range	Mag	Pha	Mag	Pha
Cao	0.0 - 0.3	**0.871**	**0.828**	**0.855**	0.743
	0.3 - 0.6	0.948	**0.957**	0.902	**0.872**
	0.6 - 1.0	0.912	**0.943**	0.885	0.903
STRM	0.0 - 0.3	0.853	0.808	0.830	**0.784**
	0.3 - 0.6	**0.949**	0.867	**0.948**	0.852
	0.6 - 1.0	**0.968**	0.922	**0.960**	**0.918**

TABLE V

DETECTION ACCURACIES OF CAO AND STRM ARE COMPARED WHEN ME1 = EPZS ≠ ME2 = DIA AND ME1 = DIA ≠ ME2 = EPZS

		EPZS, DIA		DIA, EPZS	
	Payload Range	Mag	Pha	Mag	Pha
Cao	0.0 - 0.3	0.661	0.636	0.587	0.587
	0.3 - 0.6	0.719	0.760	0.626	0.709
	0.6 - 1.0	0.756	0.719	0.725	0.654
STRM	0.0 - 0.3	**0.853**	**0.808**	**0.830**	**0.784**
	0.3 - 0.6	**0.949**	**0.867**	**0.948**	**0.852**
	0.6 - 1.0	**0.968**	**0.922**	**0.960**	**0.918**

TABLE VI

DETECTION ACCURACIES OF CAO AND STRM ARE COMPARED WHEN ME1 = HEX ≠ ME2 = ESA AND ME1 = ESA ≠ ME2 = HEX

		HEX, ESA		ESA, HEX	
	Payload Range	Mag	Pha	Mag	Pha
Cao	0.0 - 0.3	0.585	0.533	0.530	0.506
	0.3 - 0.6	0.617	0.648	0.619	0.647
	0.6 - 1.0	0.755	0.746	0.656	0.708
STRM	0.0 - 0.3	**0.765**	**0.756**	**0.842**	**0.783**
	0.3 - 0.6	**0.915**	**0.861**	**0.933**	**0.914**
	0.6 - 1.0	**0.955**	**0.914**	**0.970**	**0.975**

and re-compressed with the second motion estimation method (ME2). Stego methods employed are seperated into two groups, i.e., Phase Modifying (Pha) and Magnitude Modifying (Mag). Xu, Aly belongs to the group Mag and Pan, He, Fang belongs to the group Pha. The tests are carried out for three levels of payload ranges. Table IV,V,VI shows the test results when ME1 = ME2 and when ME1 ≠ $ME2$. The top row shows the motion estimation used during message embedding and the motion estimation used in re-compression stage. As our STRM algorithm does not requires any re-compression, the second ME (ME2) is only relevant to rows belonging to Cao *et al.*'s method.

Table IV shows that Cao *et al.* has better performance than STRM in low payloads and especially against phase modifying stego methods when ME1 = ME2. However, as it is shown in Table V and Table VI, when ME1 ≠ ME2 the accuracy of Cao *et al.* drastically drops. These tests show that Cao *et al.*'s steganalysis method is not practical when the ME1 is unknown. Moreover, STRM still has the best detection accuracy when various motion estimation methods are used.

There is a slight increase in the accuracy of STRM when ME1 = ESA. The reason is that MV is the most precise one standing for the true motion magnitude and phase. Thus, neighbouring MVs are more likely to point to the same direction and with the same strength. According to our tests, the best practice for a steganographer is not to use the full search in ME stage of the compression.

E. Choosing the Best Scheme for Motion Vector Steganalysis

In this subsection, the results of several tests with various settings which are carried out in order to find the best settings for STRM against MV steganography are presented. The same video data set in Section IV are used in the tests. The system is trained with 3 different settings:

1) Quantization factor $q = 1, 1.5, 2$ and co-occurrence distance $T = 2$,
2) $q = 1$ and $T = 2$,
3) $q = 1$ and $T = 1$,

The average detection accuracy of the system for first, second and third settings are noted as $A_{q=(1,1.5,2)}$, $A_{q=(1)}$ or $A_{T=2}$, $A_{T=1}$ respectively. It is revealed that using various quantization scales does not improve the accuracy even it does for images in [6]. The Table I exhibits the amount of improvements by choosing different schemes. The amount of increase in accuracy by choosing $q = (1, 1.5, 2)$ over $q = 1$ is listed in this table. It shows that the improvement is rather insignificant. Indeed, the accuracy falls by significantly for payload ranges (0.4-0.5, 0.5-0.6, 0.8-0.9). Therefore, the quantisation factor, q is fixed to 1 in our method. This setting also reduces the feature vector dimension to a great extent because feature array does not include the features calculated when $q = 1.5$ or $q = 2$. Feature vector size of STRM is reduced from 104013 to 44785 when only one quantization is used. The co-occurrence distance T is another variable that is tested. The feature vector size is trimmed enormously by reducing T to 1. However, test results in Table I show that contribution of larger T is considerably high. Conversely setting $T > 2$ gives a feature vector size in millions, which is practically impossible to train the classifier. Hence, we set it to $T = 2$. After combining spatial, HT and VT features with these settings the final feature vector dimension of STRM becomes 44785.

F. Why STRM Is More Accurate and Possible Shortcomings of the Proposed Method

There are five major reasons for why STRM has better performance:

1) *Utilizing spatio-temporal dependency*: It is shown that there is a temporal dependency in MV patterns which is as strong as spatial dependency (See Fig. 2). Nevertheless, there are only limited number of MV steganalysis methods take temporal dependency into account [9], [13], [15], [28], [29].
2) *Filters in higher orders*: The previous methods investigate the correlation between frames which are no further than 2 frames from each other. The filters in the proposed STRM method covers up to 5 frame distance. Larger size filters give larger coverage both spatially and temporally. In this way, STRM can exploit the correlation between any frames in two previous and two next frame range.

3) *Diverse set of filters*: We make use of the filters employed in [6]. There are six classes of filters designed for detecting different embedding distortions such as edge discontinuities, distortions in smooth regions, distortions introduced by HUGO [38] like stego algorithms etc. These diverse set of filters provide a wide spectrum attack. Assembly of many weak features are more useful than few number of strong features because the strong features are only efficient against a certain type of stego methods. As it is shown in Fig. 5, the filters designed for LSB embedding suffers when they are tested against phase modifying stego methods.
4) *Importing an already confirmed method*: Despite there is a meticulous research is going on image steganalysis, MV steganalysis methods are not inspired by the advancements in image side.[4] We have tested many image steganalysis methods on MV patterns. We found that SRM provides a competitive detection accuracy without any modification. SRM [6] and its variations [7] have been already proven to be accurate against many image steganography methods. However, it is first time that SRM is adopted to detect MV stego methods. We have modified the SRM algorithm in order to benefit from temporal dependency as well. We have also showed that image steganalysis methods could reveal the suspicious corruptions on MV patterns as well as images by a convenient adaptation.
5) *Not targeting a specific embedding algorithm*: All of the MV steganalysis methods assume that the embedding distortion would change the MV by one pixel. However, there are also phase modifying MV stego methods [11], [12], [22], [23]. We showed that the steganalysis methods could not detect these stego methods as accurate as they do LSB based stego methods. STRM does not target either LSB or phase modifying or MV perturbing [14] steganography. The resulting diverse set of rich features can capture any aberrations in MV patterns caused by message embedding.

On the other hand, memory requirement of the proposed system is relatively high. Depending on the number of the videos used in the test, training the data of size 44785×#MV planes could be infeasible for a typical personal computer. For example, in our implementation, the stego data set corresponding to 0.0-0.1 payload range and embedded with Aly's steganography required 3.2905 GB of the memory. However, this problem can be alleviated by a smarter implementation. Ensemble classifier performs training and cross validation tests on smaller dimensions. Therefore, only one chunk of the data set can be read from the storage device at a time when required by the ensemble classifier.

V. Conclusion

In this paper, we proposed a novel MV steganlaysis method by forming a rich model which is a result of many diverse high-pass filters. The filters can capture different types of dependencies among MVs in a wide spatio-temporal range. In this way, a more descriptive model of MV patterns is constructed. However, the feature vector size was too large for training with SVM. Instead, we employed ensemble classifier which is commonly used along side rich model.

We also showed that there is a strong correlation between temporally and spatially neighbouring MVs. By using this fact, we introduced a method to allow us to apply the filter to both spatial and temporal MV domain. Thus, it could capture various spatial and temporal dependencies in a longer range. Test results showed that incorporating the temporal dependency increased the detection accuracy of STRM by around 20% in low payload ranges and 5% in high payload ranges. Moreover, the proposed algorithm surpassed the previous methods in terms of classification accuracy in almost any payload.

Also, our system does not require re-compression of the stego video. The resulting steganalytic system outperformed the previous re-compression based methods as well. It has been demonstrated that the re-compression based MV steganalysis system fails when a different MV search algorithm is used in the second compression stage.

We have tested our MV steganalysis against 5 stego and 7 steganalysis methods. These large number of implementations make it the most comprehensive MV steganalysis test section in the literature. That provided a clear comparison of advantages and disadvantages of MV steganalysis and steganography algorithms took place in the tests.

Test results also showed that phase modifying stego methods are more secure than LSB based ones unless MV region is divided into insufficient number of regions. In addition to the MV modification algorithm, motion estimation method also effects the security of the stego video. The tests showed that when motion estimation used in re-compression is not same as the previously used ME, the methods based on reconstruction of MVs fail. However, our STRM algorithm performs as accurate as for any ME. Especially when full search is used, the performance of the proposed STRM method attains its maximum.

In the light of these findings, we can also add that a steganographer should avoid these three approaches: using full (exhaustive) search in motion estimation stage, employing LSB rather than phase modification and dividing regions into insufficient number of subregions. Otherwise, current MV steganography methods are readily detected by adversary methods in mid-low and higher payloads.

[4]Only exception to this is [9] and [13] are inspired by [39]

References

[1] J. Fridrich, *Steganography in Digital Media: Principles, Algorithms, and Applications*, 1st ed. New York, NY, USA: Cambridge Univ. Press, 2009.
[2] H. T. Sencar and N. Memon, *Digital Image Forensics: There is More to a Picture than Meets the Eye*, 1st ed. New York, NY, USA: Springer, 2013.
[3] R. Böhme, *Advanced Statistical Steganalysis* (Information Security and Cryptography), 1st ed. Berlin, Germany: Springer, 2010.
[4] (Feb. 2013). *Online Video: A Statistical Review*. [Online]. Available: https://www.comscore.com/Request/Presentations/2013/Online-Video-A-Statistical-Review
[5] (Sep. 2014). *Youtube Statistics*. [Online]. Available: https://www.youtube.com/yt/press/en-GB/statistics.html
[6] J. Fridrich and J. Kodovský, "Rich models for steganalysis of digital images," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 3, pp. 868–882, Jun. 2012.

[7] J. Kodovský and J. Fridrich, "Steganalysis of JPEG images using rich models," *Proc. SPIE*, vol. 8303, pp. 83030A-1–83030A-13, Feb. 2012.
[8] Y. Deng, Y. Wu, and L. Zhou, "Digital video steganalysis using motion vector recovery-based features," *Appl. Opt.*, vol. 51, no. 20, pp. 4667–4677, Jul. 2012.
[9] Y. Deng, Y. Wu, H. Duan, and L. Zhou, "Digital video steganalysis based on motion vector statistical characteristics," *Optik-Int. J. Light Electron Opt.*, vol. 124, no. 14, pp. 1705–1710, 2013.
[10] C. Xu, X. Ping, and T. Zhang, "Steganography in compressed video stream," in *Proc. 1st Int. Conf. Innov. Comput., Inf. Control (ICICIC)*, vol. 1. Aug./Sep. 2006, pp. 269–272.
[11] D.-Y. Fang and L.-W. Chang, "Data hiding for digital video with phase of motion vector," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2006, p. 4.
[12] X. He and Z. Luo, "A novel steganographic algorithm based on the motion vector phase," in *Proc. Int. Conf. Comput. Sci. Softw. Eng.*, vol. 3. Dec. 2008, pp. 822–825.
[13] C. Zhang, Y. Su, and C. Zhang, "A new video steganalysis algorithm against motion vector steganography," in *Proc. 4th Int. Conf. Wireless Commun., Netw. Mobile Comput. (WiCOM)*, Oct. 2008, pp. 1–4.
[14] Y. Cao, X. Zhao, D. Feng, and R. Sheng, "Video steganography with perturbed motion estimation," in *Information Hiding* (Lecture Notes in Computer Science), vol. 6958, T. Filler, T. Pevný, S. Craver, and A. Ker, Eds. Berlin, Germany: Springer, 2011, pp. 193–207.
[15] Y. Su, C. Zhang, and C. Zhang, "A video steganalytic algorithm against motion-vector-based steganography," *Signal Process.*, vol. 91, no. 8, pp. 1901–1909, 2011.
[16] Y. Cao, X. Zhao, and D. Feng, "Video steganalysis exploiting motion vector reversion-based features," *IEEE Signal Process. Lett.*, vol. 19, no. 1, pp. 35–38, Jan. 2012.
[17] K. Wang, H. Zhao, and H. Wang, "Video steganalysis against motion vector-based steganography by adding or subtracting one motion vector value," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 5, pp. 741–751, May 2014.
[18] K. Tasdemir, F. Kurugollu, and S. Sezer, "Video steganalysis of LSB based motion vector steganography," in *Proc. 4th Eur. Workshop Vis. Inf. Process. (EUVIP)*, Jun. 2013, pp. 260–264.
[19] K. Tasdemir, F. Kurugollu, and S. Sezer, "Spatio-temporal rich model for motion vector steganalysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 1717–1721.
[20] J. Zhang, J. Li, and L. Zhang, "Video watermark technique in motion vector," in *Proc. 14th Brazilian Symp. Comput. Graph. Image Process.*, Oct. 2001, pp. 179–182.
[21] B. Yann, L. Nathalie, and D. Jean-Luc, "A scrambling method based on disturbance of motion vector," in *Proc. 10th ACM Int. Conf. Multimedia (MULTIMEDIA)*. New York, NY, USA, 2002, pp. 89–90.
[22] P. Wang, Z. Zheng, and J. Ying, "A novel video watermark technique in motion vectors," in *Proc. Int. Conf. Audio, Language Image Process. (ICALIP)*, Jul. 2008, pp. 1555–1559.
[23] F. Pan, L. Xiang, X.-Y. Yang, and Y. Guo, "Video steganography using motion vector and linear block codes," in *Proc. IEEE Int. Conf. Softw. Eng. Service Sci. (ICSESS)*, Jul. 2010, pp. 592–595.
[24] H. A. Aly, "Data hiding in motion vectors of compressed video based on their associated prediction error," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 1, pp. 14–18, Mar. 2011.
[25] W. Jue, Z. Min-qing, and S. Juan-li, "Video steganography using motion vector components," in *Proc. IEEE 3rd Int. Conf. Commun. Softw. Netw. (ICCSN)*, May 2011, pp. 500–503.
[26] P. DeepthiChandan and M. Narayana, "High security video steganography," *Int. J. Sci. Eng. Res.*, vol. 5, no. 7, pp. 164–169, Jul. 2014.
[27] H. Zhang, Y. Cao, X. Zhao, W. Zhang, and N. Yu, "Video steganography with perturbed macroblock partition," in *Proc. 2nd ACM Workshop Inf. Hiding Multimedia Secur.* New York, NY, USA, 2014, pp. 115–122. [Online]. Available: http://doi.acm.org/10.1145/2600918.2600936
[28] H. Ye, W. Zhang, Y. Yao, C. Kong, H. Huang, and N. Yu, "Motion vector-based video steganalysis using spatial-temporal correlation," in *Proc. 6th Int. Congr. Image Signal Process. (CISP)*, vol. 1. Dec. 2013, pp. 148–153.
[29] C. Zhang, Y. Su, and C. Zhang, "Video steganalysis based on aliasing detection," *Electron. Lett.*, vol. 44, no. 13, pp. 801–803, Jun. 2008.
[30] (2014). *MPEG1/2 Standard Reference Software*. [Online]. Available: http://www.mpeg.org/MSSG/
[31] (2014). *Video Test Media [Derf's Collection]*. [Online]. Available: https://media.xiph.org/video/derf/
[32] J. Kodovský and J. Fridrich, "Steganalysis in high dimensions: Fusing classifiers built on random subspaces," *Proc. SPIE*, vol. 7880, p. 78800L, Feb. 2011.
[33] J. Kodovský, J. Fridrich, and V. Holub, "Ensemble classifiers for steganalysis of digital media," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 2, pp. 432–444, Apr. 2012.
[34] A. M. Tourapis, "Enhanced predictive zonal search for single and multiple frame motion estimation," *Proc. SPIE*, vol. 4671, pp. 1069–1079, Jan. 2002.
[35] S. Zhu and K.-K. Ma, "A new diamond search algorithm for fast block-matching motion estimation," *IEEE Trans. Image Process.*, vol. 9, no. 2, pp. 287–290, Feb. 2000.
[36] C. Zhu, X. Lin, and L.-P. Chau, "Hexagon-based search pattern for fast block motion estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 5, pp. 349–355, May 2002.
[37] (2015). *FFmpeg Library Source Code v2.4.3*. [Online]. Available: https://github.com/FFmpeg/FFmpeg/releases/tag/n2.4.3
[38] J. Kodovsky, J. Fridrich, and V. Holub, "On dangers of over-training steganography to incomplete cover model," in *Proc. 13th ACM Multimedia Workshop Multimedia Secur.*, New York, NY, USA, 2011, pp. 69–76. [Online]. Available: http://doi.acm.org/10.1145/2037252.2037266
[39] A. D. Ker, "Steganalysis of LSB matching in grayscale images," *IEEE Signal Process. Lett.*, vol. 12, no. 6, pp. 441–444, Jun. 2005.

Kasim Tasdemir received the B.Sc. degrees in electrical electronics engineering and computer engineering from Anadolu University, Eskisehir, Turkey, in 2007, the M.Sc. degree from Bilkent University, Ankara, Turkey, in 2009, and the Ph.D. degree from Queens University Belfast, U.K., in 2015. He is currently an Assistant Professor with the Department of Computer Engineering, Abdullah Gül University, Kayseri, Turkey. His research interest spans the areas of statistical signal, image and video processing, steganography, and steganalysis.

Fatih Kurugollu (M'02–SM'08) received the B.Sc., M.Sc., and Ph.D. degrees from Istanbul Technical University, Istanbul, Turkey, in 1989, 1994, and 2000, respectively, all in computer engineering. From 1991 to 2000, he was a Research Fellow with the Marmara Research Centre, Kocaeli, Turkey. In 2000, he joined the School of Computer Science, Queen's University Belfast, Belfast, U.K., as a Post-Doctoral Research Assistant. He was appointed as a Lecturer with Queen's University Belfast in 2003 and was promoted to Senior Lecturer in 2011. His research interests include multimedia security, soft computing for image and video segmentation, biometrics, and hardware architectures for image and video applications.

Sakir Sezer (M'00) received the Dipl.-Ing. degree in electrical and electronics engineering from RWTH Aachen University, Aachen, Germany, and the Ph.D. degree from Queens University Belfast, Belfast, U.K., in 1999. He is currently a Research Director and the Head of Network and Cyber Security Research with the Centre for Secure Information Technologies and holds the Chair for Secure Information Technologies with Queen's University Belfast. He is also the Co-Founder and the CTO of Titan IC Systems. He has co-authored over 160 conference and journal papers in the areas of network security, content processing, malware detection, and system-on-chip. His research is leading major (patented) advances in the field of high-performance content processing and is currently commercialized by Titan IC Systems. He is a member of the IEEE International System-on-Chip Conference Executive Committee. He was a recipient of a number of prestigious awards, including the InvestNI, the Enterprise Ireland and the Intertrade Ireland innovation and enterprise awards, and the InvestNI Enterprise Fellowship.